記憶の真髄：Titansアーキテクチャへの序章

「記憶の真の技術とは、注意の技術である！」

— サミュエル・ジョンソン, 1787年

現代AIアーキテクチャが直面する「スケーラビリティのジレマ」

完璧だが短期的な記憶

Transformerは文脈内の全トークンの依存関係を正確にモデル化するが、計算コストが文脈長に対して二乗で増加するため、扱える長さに厳しい制約がある。

無限だが損失の多い記憶

回帰型モデルは情報を固定サイズの隠れ状態に圧縮することで、効率的に長いシーケンスを処理できる。しかし、この圧縮プロセスにおいて重要な情報が失われるボトルネックを抱えている。

現在のAI分野は、「完璧だが短い文脈」か「無限だが損失の多い文脈」か、という根本的な二者択一を迫られている。

人間の記憶システム　システムに学ぶ：問題の再定義

人間の記憶は単一のプロセスではなく、短期記憶、長期記憶など、それぞれが異なる機能を持ち、独立して動作可能なシステムの連合体である。この知見は、AIアーキテクチャ設計に新たな問いを投げかける。

Q1: 優れた記憶構造とは何か？
Q2: 適切な記憶更新メカニズムとは？
Q3: 優れた記憶検索プロセスとは？
Q4: 複数の記憶モジュールをどう効率的に統合するか？
Q5: 長い過去を効果的に記憶するために「深い」記憶は必要か？

新たな記憶哲学：Titansアーキテクチャの登場

Titansは、人間の記憶システムに着想を得て、それぞれが異なる役割を担う3つのモジュールを統合することで、既存のトレードオフを乗り越える。

- **1. コア (Core / 短期記憶)**
  役割: 限定されたウィンドウサイズでの高精度な注意機構。現在の文脈を処理する。
- **2. ニューラル長期記憶 (Neural Long-Term Memory)**
  役割: 過去の文脈の抽象化された情報を学習し、パラメータ内に保存する。本アーキテクチャの核心。
- **3. 永続的記憶 (Persistent Memory)**
  役割: タスクに関する知識を符号化する、学習可能だがデータ非依存のパラメータ群。

核心技術：ニューラル長期記憶の動作原理

ニューラル長期記憶は、テスト時においても継続的に学習し、記憶を形成・管理するメタモデルとして機能する。
その動作は4つの主要な原則に基づいている。

「驚き」からの学習
(Learning from 'Surprise')

→

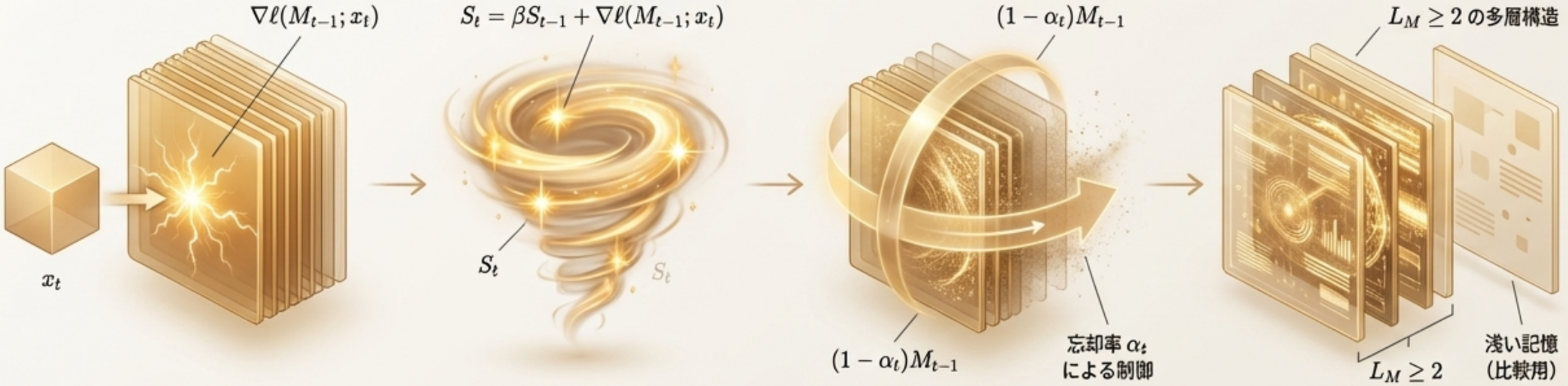
文脈の流れを捉える勢い
(Momentum for Context)

→

能動的な忘却
(Active Forgetting)

→

深い記憶アーキテクチャ
(Deep Memory Architecture)



「驚き」からの学習
(Learning from 'Surprise')

データ入力が既存の記憶モデルと相互作用し、予測誤差に基づく「驚き」の勾配が生成される。

文脈の流れを捉える勢い
(Momentum for Context)

現在の「驚き」と過去の「驚き」の蓄積を勢い項として組み合わせ、長期的な文脈を捉える。

能動的な忘却
(Active Forgetting)

古いまたは関連性の低い情報を積極的に忘却し、新しい学習のための容量を確保する。

深い記憶アーキテクチャ
(Deep Memory Architecture)

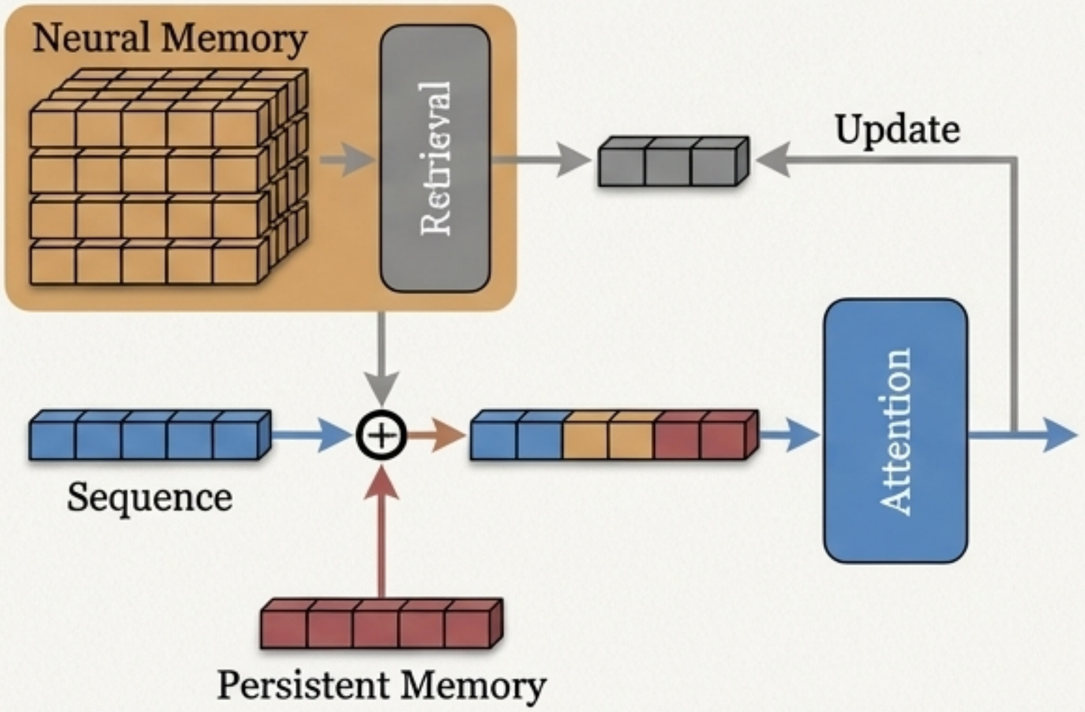
記憶構造を多層化することで、複雑で表現力の高い過去の情報を深く保持する。

3つの設計思想：Titansアーキテクチャの実装バリアント

ニューラル長期記憶は、アーキテクチャ内で多様な役割を果たすことができる。ここでは代表的な3つの実装形態を紹介する。

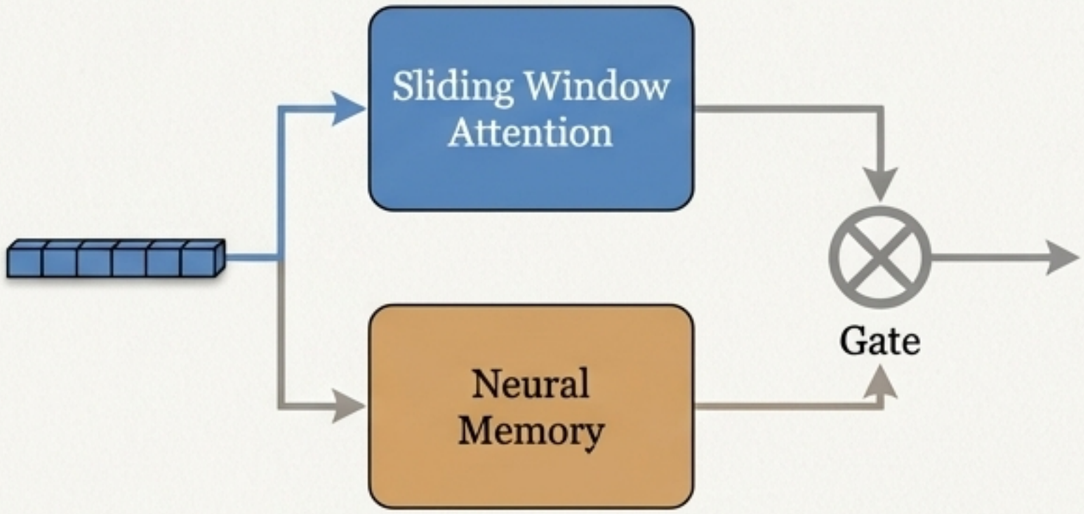
Memory as a Context (MAC)

コンセプト: 長期記憶から関連情報を検索し、現在の入力文脈と結合してアテンション層に渡す。



Memory as a Gate (MAG)

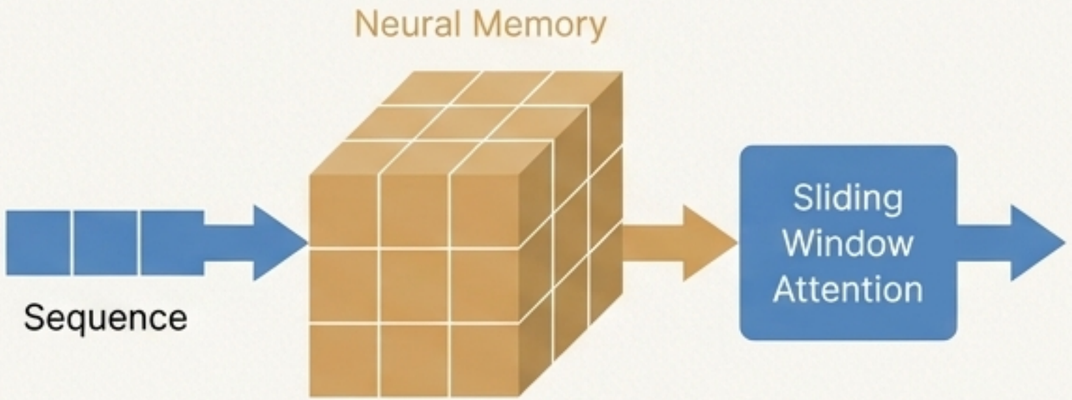
コンセプト: 短期記憶（スライディングウィンドウ・アテンション）と長期記憶（ニューラルメモリ）を並列に処理し、ゲート機構で出力を統合する。



階層的実装と効率的な並列化トレーニング

Memory as a Layer (MAL)

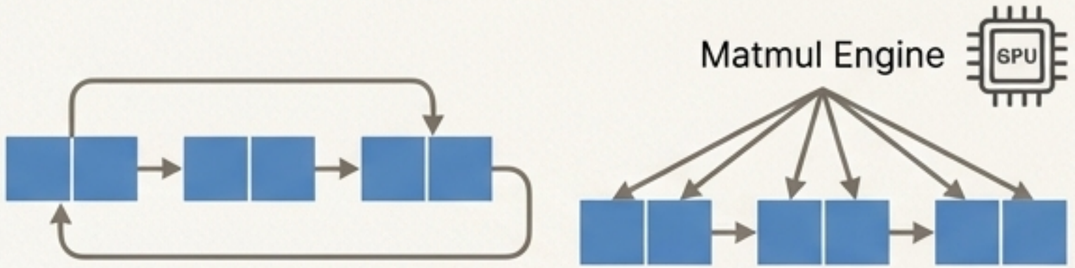
コンセプト : ニューラルメモリを一つの層として扱い、アテンション層の前段に配置する。



高速かつ並列化可能なトレーニング

課題 : 記憶の更新式は再帰的であり、逐次計算が必要に見える。

解決策 : 数式を再構成することで、更新プロセス全体を複数の行列積（Matmul）と並列スキャンで計算可能。これにより、GPU/TPUなどのハードウェアアクセラレータを最大限に活用し、高速なトレーニングを実現する。



実証：理論を裏付ける圧倒的な実験結果

Titansの有効性を検証するため、多様なタスクとモデルスケールで広範な実験を実施した。

実験設定の概要:

- **モデル:** Titansの主要バリアント（MAC, MAG, MAL, LMM）を170Mから760Mパラメータまでスケール。
- **データセット:** FineWeb-Eduからサンプリングした最大30Bトークンで学習。

評価タスク:

💬 • 言語モデリング、常識推論 🔍

🔍 • 超長文脈での情報検索（Needle in a Haystack, BABILong）

📈 • 時系列予測、DNAモデリング 🧬

主要比較対象:

- **Transformer:** Transformer++
- **最新の回帰型モデル:** Mamba, Mamba2, Gated DeltaNet, TTT

標準ベンチマークにおける卓越した性能

Titansは、言語モデリングと常識推論の両方で、同規模のTransformer++や最新の回帰型モデルを凌駕する。

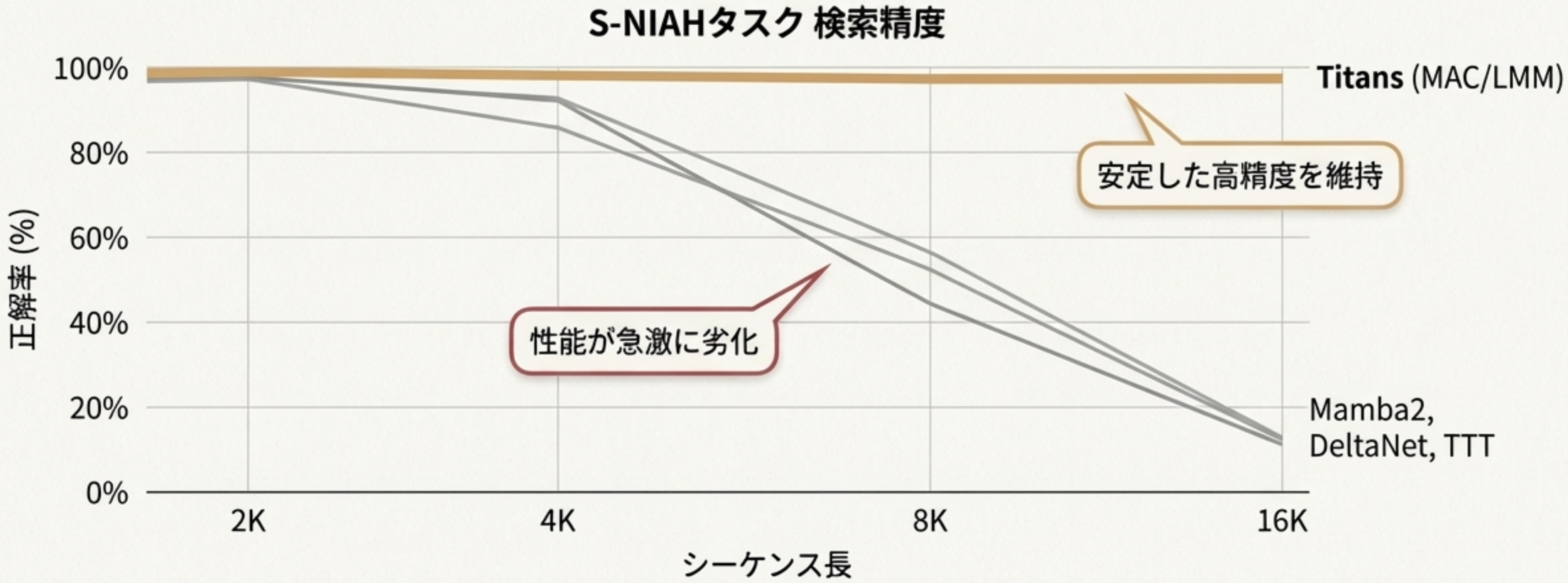
WikitextやLAMBADAにおけるパープレキシティ（低いほど良い）、およびPIQAやHellaSwagなどの常識推論タスクにおける正解率（高いほど良い）を比較。

モデル	言語モデリング（LMB. ppl↓）	常識推論（平均正解率↑）
Titans (MAG)	**19.86**	
Titans (MAC)		**52.51%**
Gated DeltaNet-H2*	20.83	51.49%
Transformer++	27.64	48.69%

真価の発揮：超長文脈での情報検索タスク (Needle in a Haystack)

文脈長が16Kトークンに及んでも、Titansは情報の検索精度を維持し、競合モデルに見られる急激な性能劣化を克服する。

RULERベンチマークのS-NIAHタスクにおける検索精度の比較。文脈が長くなるにつれて、多くのモデルは「干し草の中の針」を見つけられなくなる。

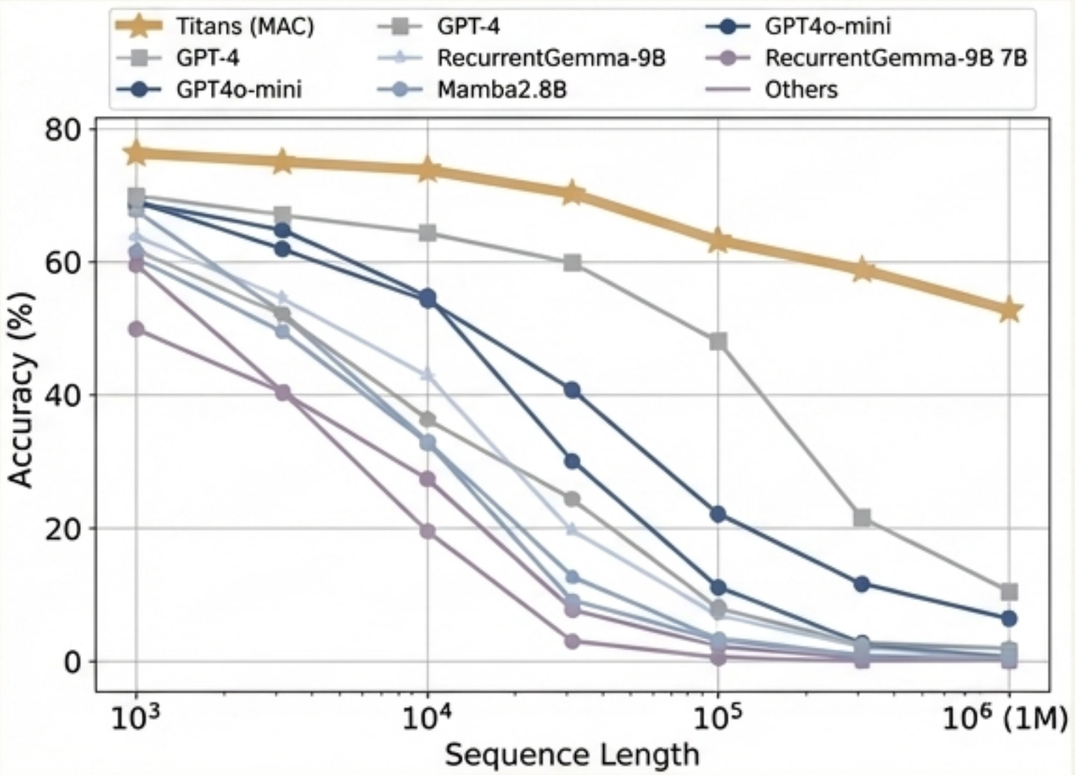


限界への挑戦：BABILongベンチマークでの驚異的な結果

BABILongは、数百万トークンに及ぶ文書全体に散らばった情報に基づく推論を要求する、最も困難な長文脈タスクの一つである。

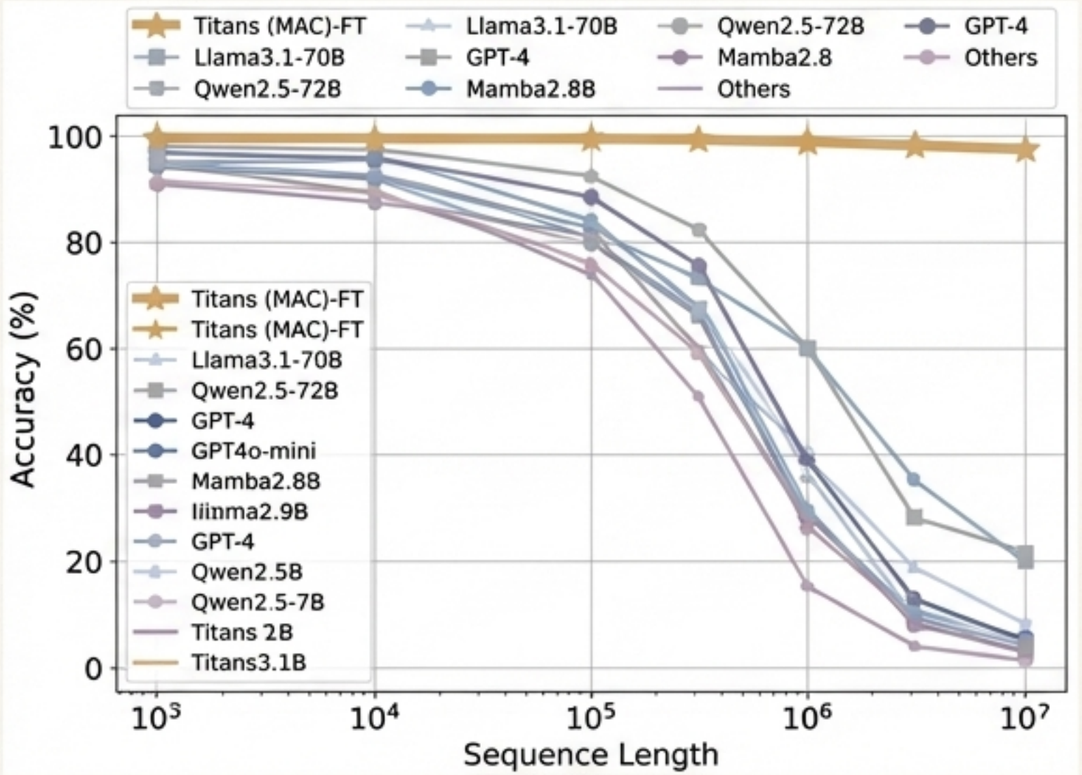
少数ショット設定（Few-shot）

Titans (MAC)は、GPT-4を含む巨大モデルを上回る性能を達成。



ファインチューニング設定（Fine-tuning）

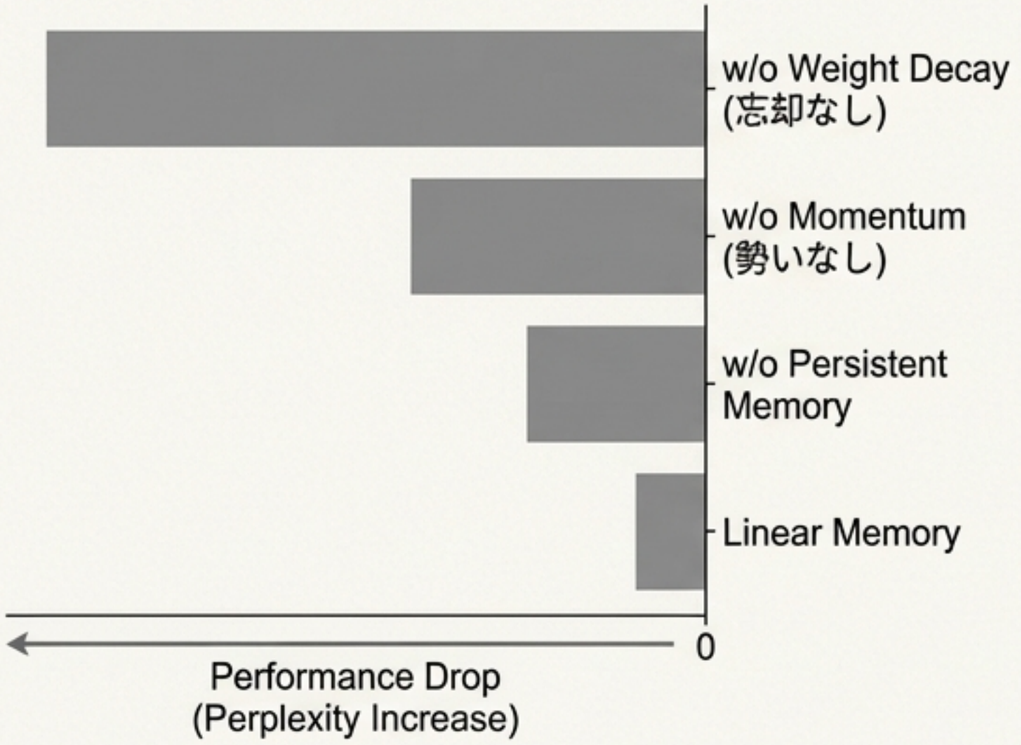
200万トークンを超える文脈においても、RAGを併用した巨大モデルや専用の長文脈モデルを凌駕。



設計の妥当性：各コンポーネントの貢献度と「深い記憶」の効果

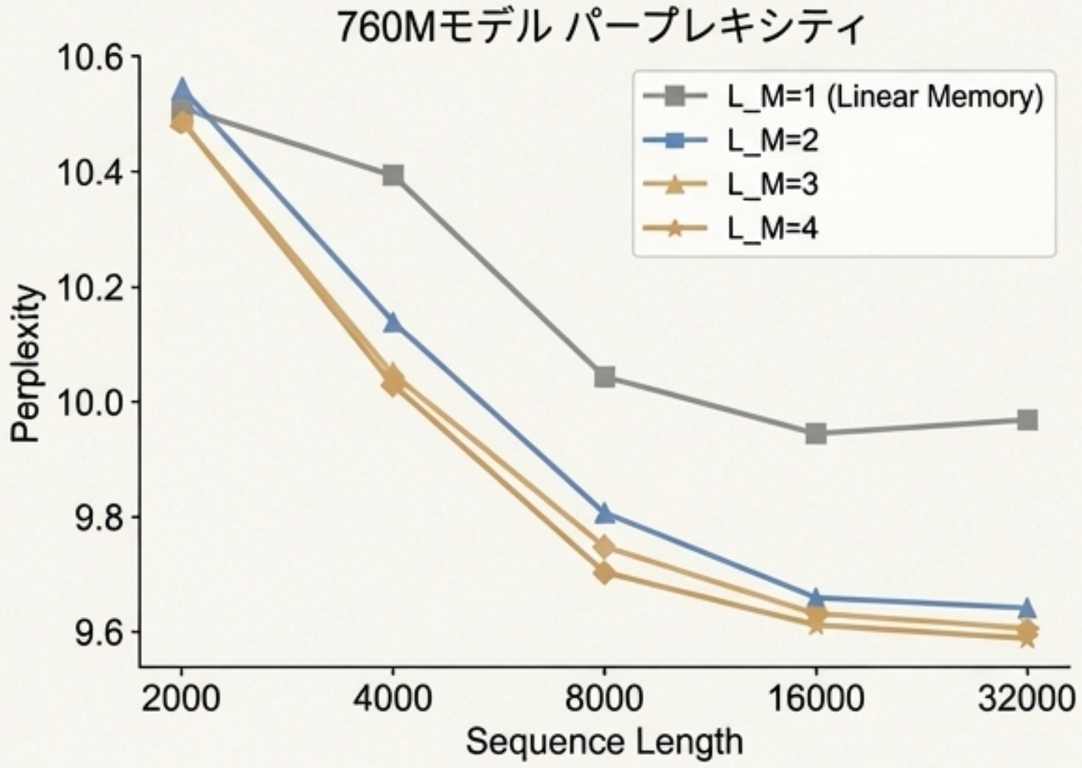
アブレーションスタディ

Titansの高性能は、全ての設計要素が不可欠な役割を果たすことで実現されている。



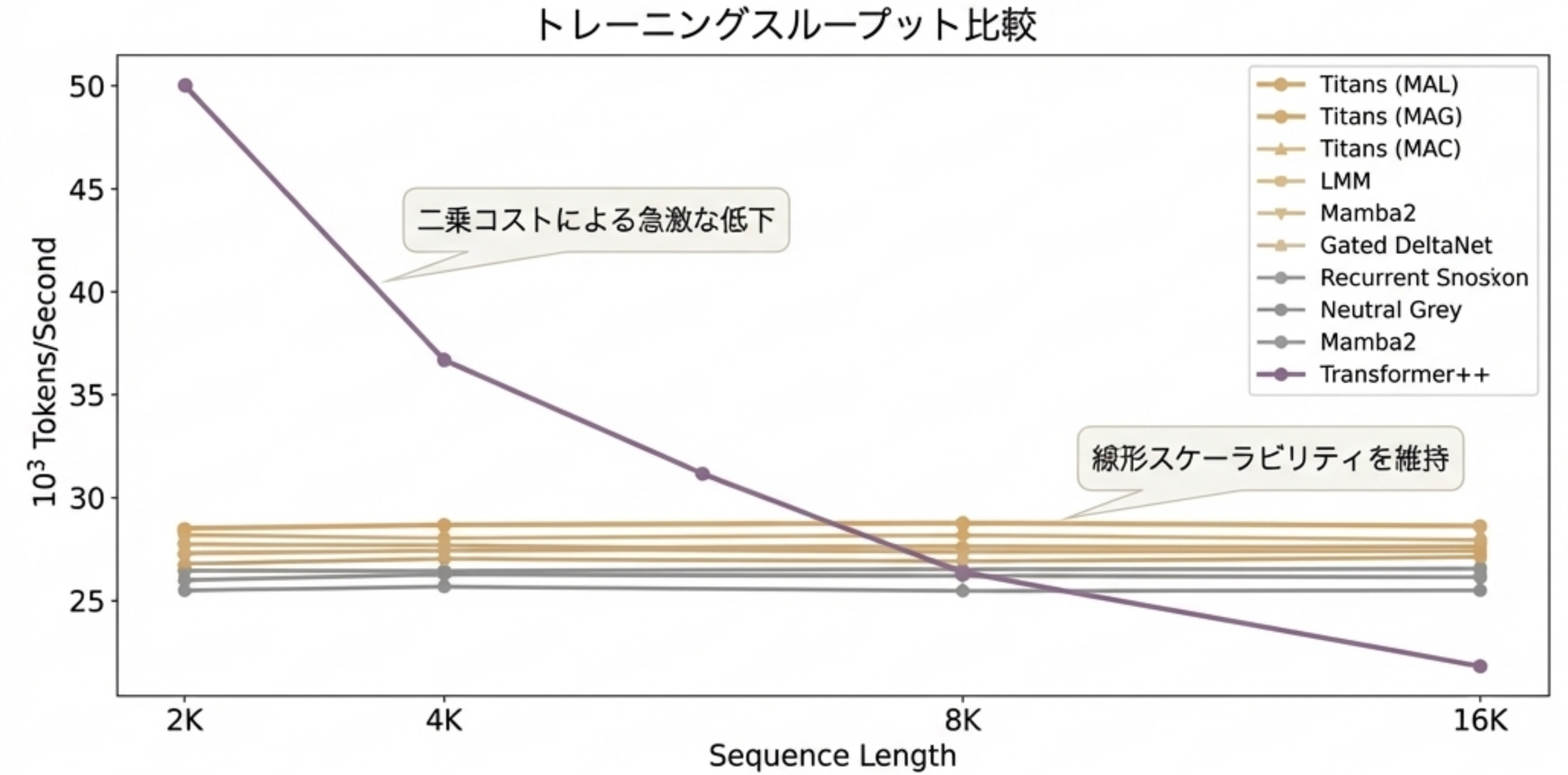
「深い記憶」の効果

記憶層が深いほど、モデルは長いシーケンスに対してより堅牢になる。



効率性とスケーラビリティ：理論と実践の両立

Titansは、Transformer++のような二乗コストのモデルに対して線形のスケーラビリティを維持し、最新の回帰型モデルに匹敵するトレーニング スループットを達成する。



自然言語の先へ：多様な領域で証明された汎用性

ニューラル長期記憶モジュール (LMM) は、その優れたシーケンスモデリング能力により、
自然言語以外のタスクでも最先端の性能を発揮する。

時系列予測（Time Series Forecasting）

- **タスク:** ETT、Traffic、Weatherなどのベンチマークで未来の数値を予測。
- **結果:** SimbaフレームワークにLMMを組み込むことで、Mambaベース、Transformerベースの既存手法を全てのデータセットで上回るMSE/MAEを達成（Table 3）。

DNAモデリング（DNA Modeling）

- **タスク:** GenomicsBenchmarksにおけるDNA配列の分類タスク。
- **結果:** HyenaDNAやMamba-Basedといったゲノム解析の最先端モデルと競合する、あるいはそれを上回る分類精度を達成（Table 4）。

記憶を再定義する：Titansが拓く次世代AIの可能性

本研究は、AIにおける「記憶」のあり方を再考し、新たなアーキテクチャファミリーを提案した。

Titansの核心的貢献：

1. **新しいニューラル長期記憶**: 認知科学に着想を得て、テスト時にも学習するメタ学習器としての記憶モジュールを設計。
2. **Titansアーキテクチャ** : 短期記憶と長期記憶を効果的に統合し、AIのスケーラビリティのジレンマを解決。
3. **最先端の性能** : 特に200万トークンを超える超長文脈タスクにおいて、既存のあらゆるモデルを凌駕する性能を実証。

結論:

Titansは、AIが情報を処理し、記憶し、推論する方法におけるパラダイムシフトを提案する。
これは、より複雑で大規模な問題を解決できる次世代AIへの重要な一歩である。